eNTERFACE'08, August 4th – August 29th, Paris, France— Project 7: Final Project Report

# Multimodal Feedback from Robots and Agents in a Storytelling Experiment

S. Al Moubayed (#), M. Baklouti (^), M. Chetouani (*), T. Dutoit (+), A. Mahdhaoui (*), J.-C. Martin (~), S. Ondas (@), C. Pelachaud (°), J. Urbain (+), M. Yilmas (&)

*(#)Center for Speech Technology, Royal Institute of Technology, KTH, SWEDEN, (^) Thalès, FRANCE, (*) University of Paris VI – FRANCE,* (+) *Faculté Polytechnique de Mons – BELGIUM, (~) LIMSI – FRANCE,* (@) Technical University of Košice – *SLOVAKIA,(°) INRIA – FRANCE, (&) Koc University – TURKEY*

***Abstract* — In this project, which lies at the intersection between Human-Robot Interaction (HRI) and Human-Computer Interaction (HCI), we have examined the design of an open-source, real-time software platform for controlling the feedback provided by an AIBO robot and/or by the GRETA Embodied Conversational Agent, when listening to a story told by a human narrator. Based on ground truth data obtained from the recording and annotation of an audio-visual storytelling database, and containing various examples of human-human storytelling, we have implemented a proof-of-concept ECA/Robot listening system. As a narrator input, our system uses face and head movement analysis, as well as speech analysis and speech recognition; it then triggers listening behaviors from the listener, using probabilistic rules based on the co-occurrence of the same input and output behaviors in the database. We have finally assessed our system in terms of the homogeneity of the database annotation, as well as regarding the perceived quality of the feedback provided by the ECA/robot.**

## I. INTRODUCTION

THIS project lies at the intersection between Human-Computer Interaction (HCI) and Human-Robot Interaction (HRI).

Human-Robot Interaction (HRI) is a multi-disciplinary field involving research on robot control (planning, sensor…), speech processing (recognition, synthesis), vision (human localization, environment characterization), artificial intelligence, cognitive science and other fields [1]. Various robots are now available for such studies, and are provided with specific programming tools. In this project, we have focused of the Sony AIBO dog, and the URBI (Universal Real-time Behaviours Interface) language [2].

Human-Computer Interaction is restricted here to Embodied Conversational Agents (ECAs). The term ECA has been coined in Cassell et al. [3] and refers to human-like virtual characters that typically engage in face-to-face communication with the human user. In this project, we have used GRETA [4], an ECA , whose interface obeys the SAIBA (Situation, Agent, Intention, Behavior, Animation) architecture [5].

Several methods have been proposed for the improvement of the interaction between humans and agents or robots. The key idea of their design is to develop agents/robots with various capabilities: establish/maintain interaction, show /perceive emotions, dialog, display communicative gesture and gaze, exhibit distinctive personality or learn/develop social capabilities [6, 7, 8]. Theses social agents or robots aim at naturally interacting with humans by the exploitation of these capabilities.

We have investigated one aspect of this social interaction: the engagement in the conversation [9]. The engagement process makes it possible to regulate the interaction between the human and the agent or the robot. This process is obviously multi-modal (verbal and non-verbal) and requires an involvement of both the partners. Some mechanisms as motivation, curiosity can be useful for this purpose [8].

Our project more specifically aims at exploring multimodal interaction between a human speaker telling a story (typically a cartoon) to (i) an ECA or (ii) an AIBO robot. More particularly, we focused on the design of an open-source, real-time software platform for designing the feedbacks provided by the robot and the humanoid during the interaction. The multimodal feedback signals we consider here are limited to facial and neck movements by the agent, while the AIBO robot uses all possible body movements, given its poor facial expressivity. We did not pay attention to arms or body gestures.

This paper is organized as follows. In Section II, we formalize SAIBA, a common architecture for embodied agents, and introduce its application to feedback modeling. This leads us, in Section III, to exposing the contents of the eNTERFACE08_STEAD database developed for this project and containing various annotated examples of

This report, as well as the accompanying database and the source code for the software developed during the project, are available online from the eNTERFACE'08 web site: www.enterface.net/enterface08.

human-human storytelling. This database was used for designing several feedback components in subsequent Sections. Sections IV and V respectively focus on the speech and video analysis modules we have used. This is followed in Section VI by details on possible control of the agent state via ASR. We then give, in Section VII, a description of the feedback rules we have established for triggering feedback from our software and hardware rendering engines, namely AIBO and GRETA, and show the behaviors we have been able to synthesize with them. Finally, Section VIII gives details on the compared performances of our HCI and HRI systems.

## II. SAIBA AND FEEDBACK

Although we are all perfectly able to provide natural feedback to a speaker telling us a story, explaining how and when you do it is a complex problem. ECAs are increasingly used in this context, to study and model human-human communication as well as for performing specific automatic communication tasks with humans).

Examples are REA [10], an early system that realizes the full action-reaction cycle of communication by interpreting multimodal user input and generating multimodal agent behaviour, the pedagogical agent Steve [11] which functions as a tutor in training situations, MAX [12] a virtual character geared towards simulating multimodal behaviour, Gandalf [13] provides real-time feedback to a human user based on acoustical and visual analysis. Carmen [14] a system that supports humans in emotionally critical situations such as advising parents of infant cancer patients. Other systems realize presentation agents [15], i.e. one or more virtual agents present some information to the user. They can adopt several roles, such as being a teacher [15; 17], a museum guide [18, 19, 12] or a companion [20, 21]. In robotics, various models have been proposed for the integration of feedbacks during interaction [7]. Recently, the importance of feedbacks for discourse adaptation has been highlighted during an interaction with BIRON [22].

In a conversation, all interactants are active. Listeners provide information to the speaker on how they view the conversation goes on. By sending acoustic or visual feedback signals, listeners show if they are paying attention, understanding or agreeing with what is being said. Taxonomies of feedbacks, based on the meaning these signals convey, have been proposed [23, 24]. The key idea of this project is to automatically detect the communicative signals in order to produce a feedback. Contrary to the approach proposed in [LOH08], we focus on non-linguistic features (prosody, prominence) but also on head features (activity, shake, nod).

Our system is based on the architecture proposed by [4], but progressively adapted to the context of a storytelling (figure 1). We developed several modules for the detection and the fusion of the communicative signals from both audio and video analysis. If these communicative signals match our pre-defined rules, a feedback is triggered by the Realtime BackChannelling module, resulting in two different behaviors conveying the same intention.

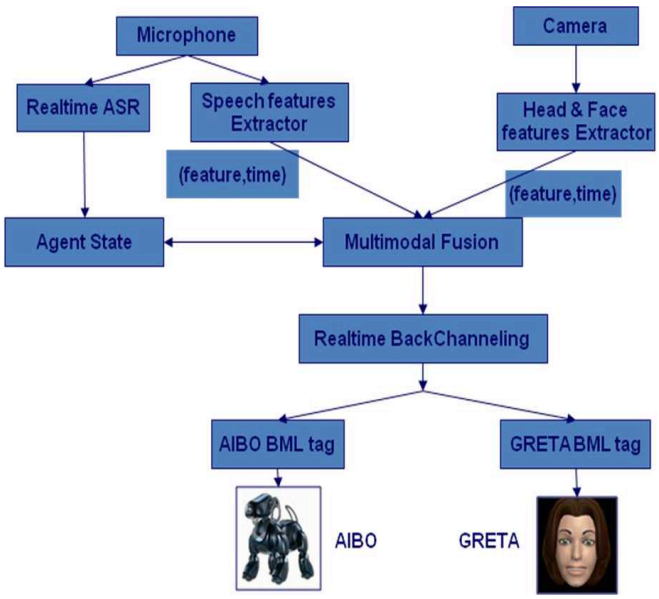
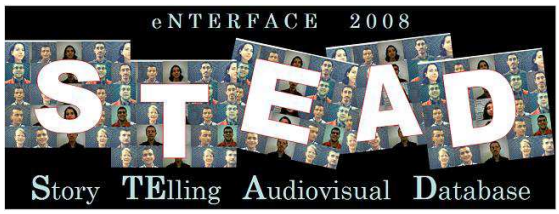



**Fig. 1**. Architecture of our interaction feedback model

## III. THE ENTERFACE08_STEAD DATABASE




In order to model the interaction between the speaker and the listener during a storytelling experiment, we first recorded and annotated a database of human-human interaction: the eNTERFACE08_STEAD database. This database was used for extracting feedback rules (section VII, but also for testing the multi-modal feature extraction system (section VIII).

We followed the McNeill lab framework [25]: one participant (the speaker), who has previously observed an animated cartoon (Sylvester and Tweety), tells the story to a listener immediately after viewing it. The narration is accompanied by spontaneous communicative signals (filled pauses, gestures, facial expressions, etc.). In contract, instructions are given to the listener to express his/her engagement in the story by giving non-verbal audio-visual gestures in response to the story told by the speaker.

### eNTERFACE_STEAD Contents

Twenty-two storytelling sessions telling the "Tweety and Sylvester - Canary row" cartoon story were recorded.

Thirteen recording sessions were done by a French listener and a French speaker. The last two recordings have exaggerated non-verbal activity (closer to acting than to real-life storytelling).

Four recording sessions were done by an Arabic listener and an Arabic speaker.

Five recording sessions were done by a speaker and a listener who do not speak or understand each other's languages; these recordings can be used to study the isolated effect of prosody on the engagement in a storytelling context. The languages used in these sessions were Arabic, Slovak, Turkish, and French.

**Annotation Schema**

A part of the MUMIN [26] multimodal coding scheme was used for annotating the database. MUMIN was originally created to experiment with annotation of multimodal communication in video clips of interviews taken from Swedish, Finnish and Danish television broadcasting and in short clips from movies. However, the coding scheme is also intended to be a general instrument for the study of gestures and facial displays in interpersonal communication, in particular the role played by multimodal expressions for feedback, turn management and sequencing.

The videos were annotated (with at least two annotators per session) for describing simple communicative signals of both speaker and listener: smile, head nod, head, shake, eye brow and acoustic prominence. These annotations were done using the ANVIL [27] annotation program (Fig. 2) and are summarized in Table 1.

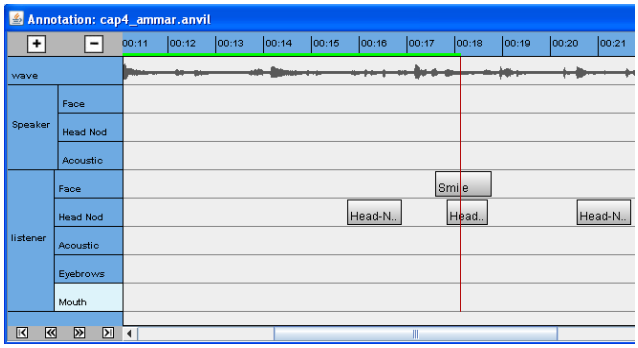



**Fig. 2** Anvil, as used for our database annotation

| **Facial display feature** | **Form of expression/Movement values** | |
|---|---|---|
| **Facial display feature** | **Value** | **Short tag** |
| **General face** | smile | smile |
| **Mouth(opening)** | Open mouth | Open-M |
| | Closed mouth | Close-M |
| **Head** | Nod | Head-Nod |
| | Shake | Head-shake |
| **Eyebrows** | Frowning | Frown |
| | Raising | Raise |
| **Acoustic** | Prominence | Prominence |
| | laughter | laughter |

**Table 1:** Coding scheme used for eNTERFACE_STEAD annotations.

Annotators had instructions to annotate prominence in the audio recording of the speaker by only listening to the audio signal without looking to the video recording; some annotators spoke French, others did not.

**Annotation evaluation**

Manual annotations of videos were evaluated by computing agreements using corrected kappa [28] computed in the Anvil tool [27], shown in Fig. 2:

$$kappa = (P0 - 1/z) / (1 - 1/z)$$

where $z$ is the number of categories and $P0$ is like in Cohen's kappa.

Table 2 presents the agreements among annotators for each track. We can see that the best agreement is obtained for the Listener.Acoustic track which is expected since the listener is not assumed to speak and when he/she does simple sounds are produced (filled pauses). Other tracks have a lower agreement such as Speaker.Acoustic. The speaker always speaks during the session and prominent events are less identifiable. However, the agreements measures are high enough to allow us to assume that selected communicative signals might be reliably detected.

| **Track Name** | **Cohen's Kappa** | **Corrected Kappa** | **Agreement(%)** |
|---|---|---|---|
| Speaker.Face | 0.473 | 0.786 | 89.306 |
| Speaker.Acoustic | 0.099 | 0.786 | 84.500 |
| Listener.Face | 0.436 | 0.559 | 77.960 |
| Listener.HeadNod | 0.464 | 0.694 | 84.622 |
| Listener.Acoustic | 0.408 | 0.929 | 95.972 |

**Table 2** Agreement between annotators of our database (eNTERFACE_STEAD)

**The eNTERFACE_STEAD License**

The eNTERFACE_STEAD contents, and all the annotations are released under an MIT-like free software license and is available from the eNTERFACE'08 website (www.enterface.net/enterface08).

## IV. Speech Analysis

The main goal of the speech analysis component is to extract features from the speech signal that have been previously identified as key moments for triggering feedbacks (cf. section VII). In this study, we do not use any linguistic information to analyze the meaning of the utterances being told by the speaker, but we focus on the prosodic cross-language features which may participate in the generation of the feedback by the listener.

Previous studies have shown that pitch movements, especially at the end of the utterances, play an important

role in turn taking and backchannelling during human dialogue [29]. In this work, we propose in this work to use the following features extracted from the speaker's speech signal: Utterance beginning, Utterance end, Raising pitch, Falling pitch, Connection pitch, and Pitch prominence.

To extract these important features from the speech stream, we decided to work in Pure Data [30], a graphical programming environment for real-time audio processing. The patch we developed for speech feature extraction is shown in figure 3; it provides the following features: Utterance beginning, Utterance end, Raising pitch, Falling pitch, Stable pitch, and Acoustic prominence.

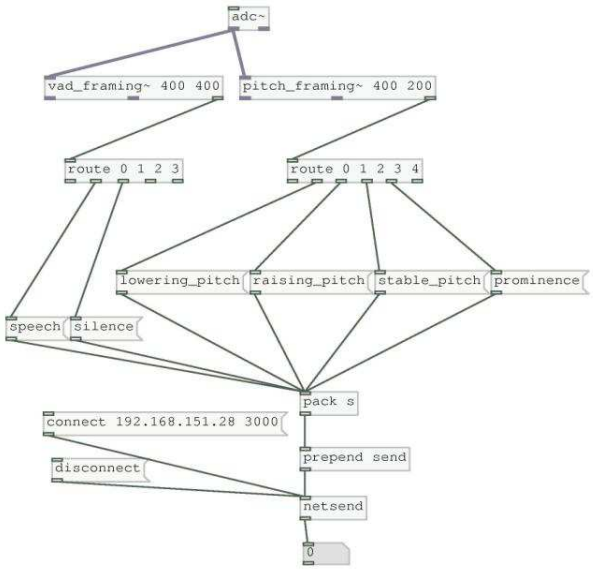



**Fig. 3**. Pure Data patch for Speech Feature Extraction

Audio acquisition is performed by the `adc~` object. It provides 64-samples blocks (at a sampling frequency set to 16kHz) to objects `vad_framing~` and `pitch_framing~`, which are responsible for voice activity detection and pitch estimation for prominence estimation. These algorithms are written in C, using audio processing functions from a C library developed by the Center of Speech Technology at KTH, Sweden. They were compiled as externals for Pure Data, so that they can be used as ordinary Pure Data objects.

Since we wanted to compute some features on overlapping audio segments longer than 64 samples, we developed a specific framing routine. The objects `vad_framing~` and `pitch_framing~` take two arguments *X* and *Y*, which impose blocks of *X* samples with a shift of *Y* samples between successive blocks: a buffer of *X* samples is filled with the input blocks of 64 samples; when this buffer is full, it is sent to the analysis algorithm, the samples in the buffer are then shifted by *Y* samples, and the buffer is filled again, etc.

For every input speech frame, `vad_framing~` and `pitch_framing~` output an integer, which is set to 0 if there is no event (feature) is detected in the speech, and to a number (index) indicating the id of the detected feature otherwise. These indices are sent to the "route" PureData object, which triggers a string depending on its input; this string is later sent through a tcp/ip connection to the Multimodal Fusion module.

The vad_framing~ object is a Voice Activity Detection object, which contains an adaptation of the SPHINX Vader functionality [31]. This object sends "1" if there is a detected change in the audio stream from Silence to Speech, and "2" when there is a detected change from Speech to Silence, otherwise the output of this object is always "0".

The pitch_framing~ object is used to extract the rest of the speech features. This object contains an implementation of the realtime fundamental frequency tracking algorithm YIN [32]. For cleaning the output of the YIN algorithm, a median filter of size "5" (60 msec) is applied on the extracted F0 to compensate for outliers and octave jumps.
This object sends "1" when a Raising Pitch is detected, "2" for Falling pitch, "3" for Stable Pitch, and "4" for Acoustic prominence.

**The TILT model:**
The TILT model [33] is used to extract Raising pitch, Falling pitch, Connection pitch. We tried in this implementation to not compensate for the unvoiced segments by using any type of interpolation; nevertheless, the movements of the pitch are detected only at the end of the voiced segments, and no movements are detected when the voiced segment duration is shorter than 125 msec.

**Audio prominence estimation**
In the literature, several definitions of acoustical prominent events can be found showing the diversity of this notion [34, 35]. Terken [35] defines prominence as words or syllables that are perceived as standing out from their environment. Most of the proposed definitions are based on linguistic and/or phonetic units.

We propose in this project another approach using statistical models for the detection of prominence. The key idea is to assume that a prominent sound stands out from the previous message. For instance, during our storytelling experiment, speakers emphasize words or syllables when they want to focus the attention of the listener on important information. These emphasized segments are assumed to stand out from the other ones, which makes them become salient.

Prominence detectors are usually based acoustic parameters (fundamental frequency, energy, duration, spectral intensity) and machine learning techniques (Gaussian Mixture Models, Conditional Random Fields) [36, 37]. Unsupervised methods have been also investigated such as the use of Kullback-Leibler (KL) divergence as a measure of discrimination between prominent and non-prominent classes [38]. These statistical methods provide an unsupervised framework adapted to our task. The KL divergence needs the estimation of two covariance matrices (Gaussian assumption):

$$KL_{ij} = \frac{1}{2}\left[ \log\frac{\Sigma_j}{\Sigma_i} + tr\left(\Sigma_i \Sigma_j^{-1}\right) + \left(\mu_i - \mu_j\right)^T \Sigma_j^{-1} \left(\mu_i - \mu_j\right) - d \right]$$

where $\mu_i$, $\mu_j$ and $\Sigma_i$, $\Sigma_j$ denote the means and the covariance matrices of *i*-th (past) and *j*-th (new event) speech segments respectively. *d* is the dimension of the speech feature vector. An event *j* is defined as prominent if the distance from the past segments (represented by the segment *i*) is larger than a pre-defined threshold.

One major drawback of the KL divergence approach is that since the new event is usually shorter than the past events, the estimation of their covariance matrices is less reliable. In addition, it is well-known that duration is an important perceptual effect for the discrimination between sounds. Taking these points into account, we propose to use another statistical test namely the $T^2$ Hotteling distance defined by:

$$H_{ij} = \frac{L_i L_j}{L_i + L_j}\left[ \left(\mu_i - \mu_j\right)^T \Sigma_{i \cup j}^{-1} \left(\mu_i - \mu_j\right) \right]$$

where $i \cup j$ is the union of *i*-th (past) and *j*-th (new event) segments. $L_i$ and $L_j$ denote the length of the segments. The $T^2$ Hotteling divergence is closely related to the Mahalanobis distance.

In this work only the fundamental frequency (*F0*) is used as a feature to calculate the Hotteling distance between two successive voiced segments. In this sense, a prominence is detected when the Hotelling distance between the current and the preceding Gaussian distributions of *F0* is higher than a threshold. We have used a decaying distance threshold over time, where the initial value of this threshold is the highest distance during the first utterance of the speaker; whenever this threshold is reached by a following segment, a Pitch Prominence event is triggered, and the new distance becomes the distance threshold. Since we estimate a Gaussian distribution of the pitch for a voiced segment, we only estimate it when there are enough pitch samples during the voiced segment, (we set this duration threshold to 175 msec).

## V. Face Analysis

The main goal of the face analysis component (Fig. 4) is to provide the feedback system with some knowledge of communicative signals conveyed by the head of the speaker. More specifically, detecting if the speaker is shaking the head, smiling or showing neutral expression are the main activity features we are interested in. The components of this module (Fig. 5) are responsible for face detection, head shake and nod detection, mouth extraction, and head activity analysis. They are detailed below.

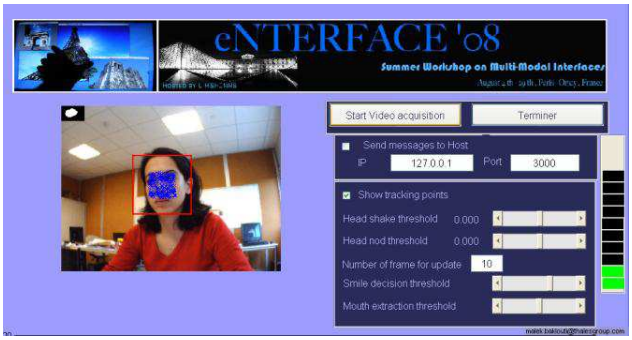



**Fig. 4**. Screenshot of the face analysis component, which runs on a PC and shows the analysis results through a MATLAB-based user interface It sends its results to the multimodal fusion module via TCP-IP.

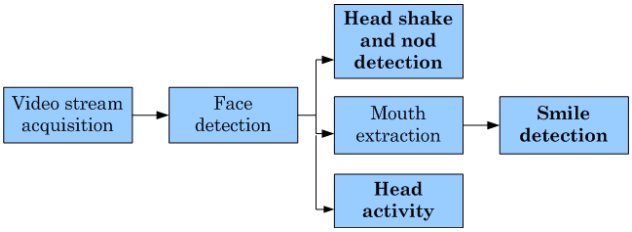



**Fig. 5**. Overview of the face analysis module

### Face detection

The face detection algorithm that we used exploits Haar-like features that have been initially proposed by Viola & Jones [39]. It is based on a cascade of boosted classifiers working with Haar-like features and trained with a few hundreds of sample views of faces. We used the trained classifier available in OpenCV.

The face fetection module outputs the coordinates of existing faces in the incoming images.

### Smile detection

Smile detection is performed in two steps: mouth extraction followed by smile detection. We use a colorimetric approach for mouth extraction. A thresholding technique is used after a colour space conversion to the YIQ space (Fig. 6).

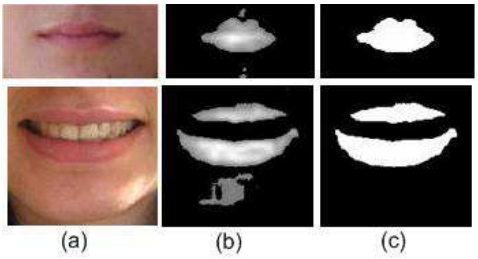



**Fig. 6**. Mouth extraction. (a) original images (b) color conversion and thresholding (c) elimination of small regions

Once the mouth is extracted, we examine the ratio between the two characteristic mouth dimensions, P1P3 and P2P4 (see Fig. 7), for smile detection. We assume that when smiling, this ratio increases. The decision is obtained by thresholding.

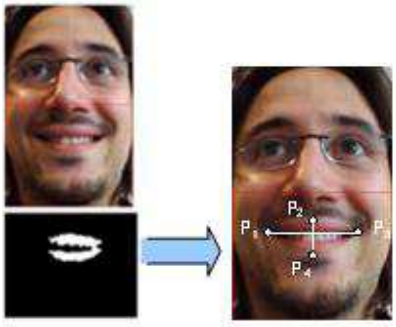



**Fig. 7**. smile detection

**Head shake and nod detection**

The purpose of this component is to detect if the person is shaking his/her head or doing a nod. The idea is to analyze the motion of some feature points extracted from the face along the vertical and horizontal axes.

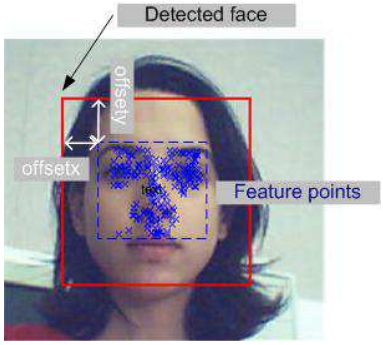



**Fig. 8**. Feature points extraction

Once the face has been detected in the image, we extract 100 feature points using a combined corner and edge detector defined by Harris [40]. Feature points are extracted in the central area of the face rectangle using offsets (Fig 8).

These points are then tracked by calculating the optical flow between a set of corresponding points in two successive frames. We make use of the Lucas Kanade [41] algorithm implementation available from in the OpenCV library (http://sourceforge.net/ projects/ opencvlibrary/).

Let $n$ be the number of feature points and $Pt_i(x_i, y_i)$ the $i$th feature point defined by its 2D screen coordinates $(x_i, y_i)$. We then define then the overall velocity of the head as:

$$V = \begin{cases} V_x = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_{i-1}) \\ V_y = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_{i-1}) \end{cases}$$

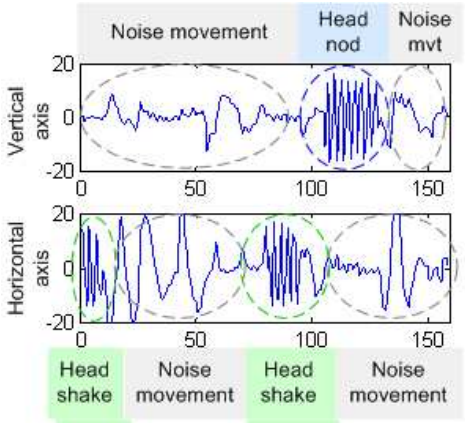



**Fig. 9**. Feature point velocity analysis

Fig. 9 shows the velocity curves along the vertical and horizontal axes. The sequence of movements represented is composed by one nod and two head shakes. We notice that the velocity curves are the sum of two signals: (1) a noise movement which is a low frequency signal representing the global head motion and (2) a high frequency signal representing the head nods and head shakes.

The idea is then to use wavelet decomposition to remove the low frequency signals. More precisely, we decomposed the signal using symlet-6 wavelet. Fig. 10 shows the reconstruction of the detail at the first level of the signal shown in Fig. 8. The head nod and shake events can be reliably identified by this process.

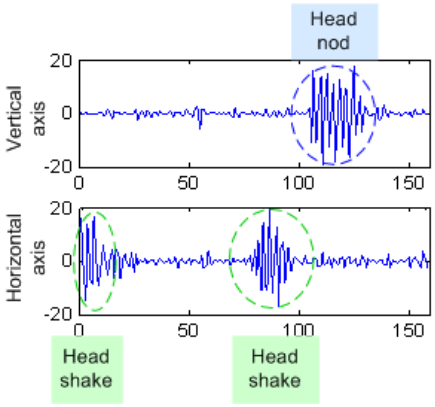



**Fig. 10** Signal denoising via wavelets.

**Head activity analysis**

Analysis of recordings of the storytelling experience has shown a correlation between the head activity of both speaker and listener. To characterize the head activity, we use the velocity of the feature points defined in (1), to quantify the overall activity A:

$$A = \sum_{t \in timewindow} V_{x,t}^{\ 2} + V_{y,t}^{\ 2} \qquad (2)$$

where *timewindow* is set to 60 frames (30 frames/s).

This measure provides information about the head activity levels. In order to quantize head activity into levels (high, medium or low), we analyzed the head activity of all the speakers of the eNTERFACE08_STEAD corpus. Assuming that the activity of one given speaker is Gaussian, we set up various thresholds defined in Table 3. By using these thresholds, the algorithm becomes more sensitive to any head movement of a stationary speaker, while it raises the thresholds for an active speaker, thus resulting in a flexible adaptive modeling.

| **Amplitude** | **Interpretation** |
|---|---|
| < mean | LOW ACTIVITY |
| < mean + standard deviation | MEDIUM ACTIVITY |
| Otherwise | HIGH ACTIVITY |

**Table 3**. Segments are categorized according to the amplitude of their maxima. Mean and standard deviation statistics are related to head activity.

## VI. AGENT/ROBOT STATE CONTROL

In the feedback model proposed in [4], the state of the agent/robot is characterized by the following features: *disagreement, agreement, acceptance, refusal, belief, disbelief, liking, disliking, interest, no_interest, understanding, no_understanding,* and *mimicry*. For this project, we have reduced this set to: *interest, understanding,* and *liking*. Our goal was then the design means of modifying these features through some analysis of the audio and video streams from the speaker.

To achieve this goal, we have used an English speaking ASR system based on keyword spotting, which makes it possible to modify the agent state according to the recognized words. The ASR system is thus integrated into an *Agent State Manager* (ASM) module (Fig 11), which consists in three main parts: the ASR engine, the State Planner and the Message Generator. These components are detailed in the next paragraphs.

**ASR engine**

The speech engine we have used is based on ATK/HTK [42] and is available as a dynamic-link library (dll), whith a simple API.

It uses freely available British English triphone acoustic models, which are part of the ATK distribution, and were trained on the WSJCAM0 speech corpus [43] recorded at the Cambridge University and composed of readings of the Wall Street Journal.

As a language model we have used a speech grammar which enables the recognition of keywords in phrases. Other words are modeled as "filler words", and not recognized. The keyword spotting grammar actually puts all keywords in parallel (with no specific syntactic constraints), together with a filler model. It is written in BNF (Backus Naur Form) format and then it is translated to HTK-compatible SLF format.

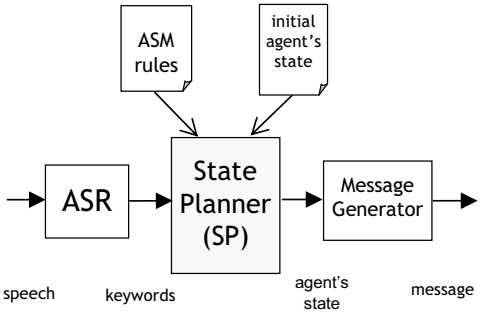



**Fig. 11**. Architecture of the Agent State Manager

For the purpose of storytelling we needed to define appropriate keywords. The best way to obtain this information was to look at our storytelling recordings, available in the from eNTERFACE08_STEAD database (see Section III). We used two recordings by native English speakers, as well as recordings in Slovak, and transcribed them into English words. The transcriptions were then analyzed in terms of word counts via a simple Python script. After eliminating the articles (*the*, *a*), conjunctions (*and*) and pronouns (*she, he, it*), which were naturally the most frequent words, we identified a group of keywords for our storytelling experiment, containing: *hotel, reception, door, room, stairs, luggage, baggage, bags, birdcage, umbrella, clerk, tomcat, cat, silvester, woman, lady, tweety, bird, knock, carry, call, run, trick, discover, hit, hide, ring, uncover, pick up, cartoon, story, treatment, outside.*

Notice that the ASR engine also uses a simple pronunciation dictionary to specify the expected pronunciation of keywords. We modified the default pronunciation of words in the dictionary by removing the expected short pauses (sp) after each word. (As a matter of fact, we use continuous speech in this project, in which there are no short pauses between words.)

**State Planner**

The State Planner is the main part of the Agent state Manager. Its task is to modify the state of the agent according to spoken input (keywords). For this purpose it uses a rule-based approach. This component is initialized with the initial agent state as well as with a set of rules loaded from an initialization file. It takes spoken keywords as input, and changes the value of the *interest, understanding,* and *liking* as output.

Each rule consists of the following fileds: *feature*, *keyword*, *step*, *max_value* and *opposite_feature* (which is optional). When the speech recognizer recognizes a *keyword*, State Planner looks for an appropriate rule. If it finds it, it increases the value of the related *feature* by a given *step* value, while *max_value* is not reached. If some *opposite_feature* is defined in the rule, it is decreased by the same step.

**Message Generator**

Every three words which have triggered the application of rules, the Message Generator prepares an XML file containing the feature values, which represent the new agent

state. After converting special characters into hexadecimal, it sends the XML file through a TCP socket to the ECA (Greta).

## VII. Multimodal fusion and Feedback Behaviors for AIBO and GRETA

**Extracting rules from data**

Based on the selected communicative signals, we have defined some rules to trigger feedbacks. The rules are based on [44, 45], which involved mainly only mono-modal signals. The structure of such rules is as follows:

"If some *signal* (eg. head-nod, pause, pitch accent) is received, then the listener sends some *feedback signal* with probability X."

We have extended these rules by analyzing the data annotated from our eNTERFACE08_STEAD storytelling database. We looked at the correlation of occurrence between each speaker mono-modal and multi-modal signal and each listener feedbacks (where we understand *multi-modal signal* as any set of overlapping signals that are emitted by the speaker within a time window, defined as the time interval of any speaker signal plus 2 seconds). This gave us a correlation matrix between speaker and listener signals, whose elements give, for each speaker signal, the probability that the listener would send a given feedback signal. In our system we use this matrix to select listener's feedback signals. When a speaker's signal is detected, we choose from the correlation matrix, the signal (ie feedback) with the higher probability.

From this process, we identified a set of rules (which can be found in the repository of the project) such as:

- Mono-modal signal ⇒ mono-modal feedback: *head_nod* is received, then the listener sends *head_nod_medium*.
- Mono-modal signal ⇒ multi-modal feedback: *smile* is received, then the listener sends *head_nod and smile.*
- Multi-modal signal ⇒ mono-modal feedback: *head_activity_high* and *pitch_prominence* are received, then the listener sends *head_nod_fast*.
- Multi-modal signal ⇒ multi-modal feedback: *pitch_prominence* and *smile* are received, then the listener sends *head_nod* and *smile*.

Rules can be made probabilistic via associated probabilities: in case there is more than one rule with the same input, every rule will have a probability of execution.

**Multi-modal fusion**

The multi-modal fusion module is responsible for activating the rules mentioned above, when input signals are detected, and will eventually trigger feedbacks from the agent/robot.

For realtime consideration, the rule contains a response time variable, which defines when the output of the rule should be executed after the reception of the last input signal; the last variable is rule duration, rule duration defines how long this rule can be active, so in case not all the input signals are received, the rule will be deactivated after this specified period.

**Reactive behaviors**

In our architecture, we aim to drive different types of virtual and/or physical agents: the GRETA ECA (Fig 12), and The AIBO ERS-7 Robot (Fig 13). To ensure high flexibility we are using the same control language to drive all the agents, the Behavior Markup Language BML [5]. BML encodes multimodal behaviors independently from the animation parameters of the agents.

Through a mapping we transform BML tags into MPEG-4 parameters for the GRETA agent and into mechanical movements for the AIBO robot. Various feedbacks are already available for GRETA such as acceptance (*head_nod*), non-acceptance (*head_shake*) or smile. Concerning AIBO, we developed similar feedbacks conveying the same meaning but in a different way. To develop the reactive behavior of AIBO, we used the URBI (Real-Time Behavior Interface) library [2] allowing a high-level control of the robot.

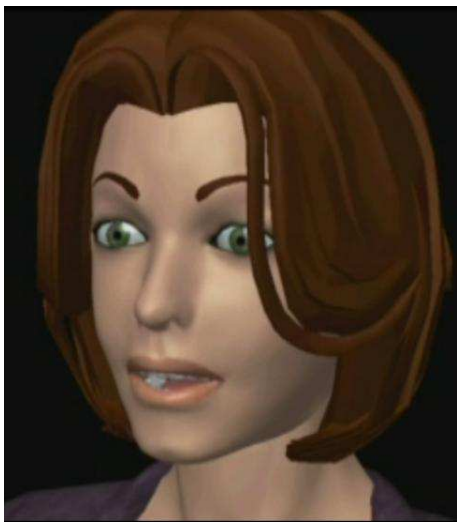


**Fig. 12** The GRETA ECA

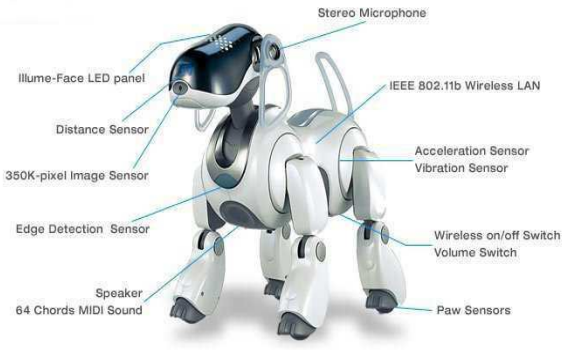



**Fig. 13** The ERS-7 Sony Aibo Robot. (from http://www.sony.net/Products/aibo/)

## VIII. Assessment and Discussion

Evaluation research is still underway for virtual characters [46, 47, 48] and for human-robot interaction [49, 50].

Since the goal of the project was to compare feedback provided by two types of embodiments (a virtual character and a robot) rather than to evaluate the multimodal feedback rules implemented in each of these systems, we decided to have users tell a story to both Greta and Aibo at the same time. The feedback system tested was the one described in the previous Sections, with the exception of the ASR system, which was not used here.

An instruction form was provided to the subject before the session. Then users watched the cartoon sequence, and were asked to tell the story to both Aibo and Greta (Fig 14). Finally, users had to answer a questionnaire. The questionnaire was designed to compare both systems with respect to the realization of feedback (general comparison between the two listeners, evaluation of feedback quality, perception of feedback signals and general comments). The evaluation form is provided in appendix. Sessions were videotaped using a Canon XM1 3CCD digital camcorder.

**Fig 14**. The assessment set-up

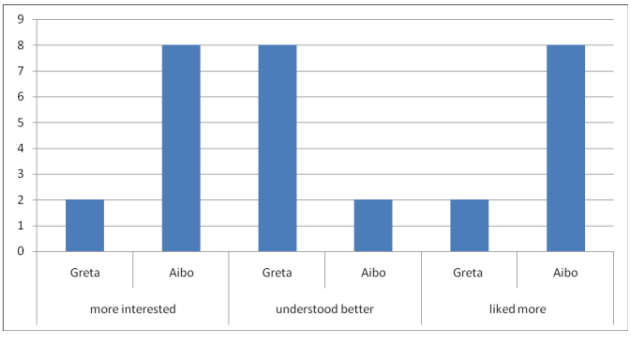



**Table 4** Comparing the feedback provided by the virtual character and the robot

As illustrated by table 4, 8 out of 10 users estimated that GRETA understood better the story than AIBO. Yet, 8 out of 10 users felt that AIBO looked more interested and liked the story more than GRETA did.

Further evaluations could be investigated with such a system. Another possibility would be to have the speaker tell two different stories one to Greta, and then another one to Aibo. The order of the listeners should be counterbalanced across subjects. This would avoid having the speaker to switch his attention between Aibo and Greta. Perceptive tests on videos combining speakers and Aibo/Greta listeners could also be designed to have subjects 1) compare random feedback with feedback generated by analyzing user's behavior , or 2) rate if the listener has been designed to listen to this speaker or not.

## IX. Conclusion and Further Developments

We presented a multi-modal framework to extract and identify Human communicative signals for the generation robot/agent feedbacks during storytelling. We exploited face-to-face interaction analysis by highlighting communicative rules. A real-time feature extraction module has been presented allowing the characterization of communicative events. These events are then interpreted by a fusion process for the generation of backchannel messages for both AIBO and GRETA. A simple evaluation was established, and results show that there is an obvious difference in the interpretation and realization of the communicative behavior between humans and agents/robots.

Our future works are devoted to the characterization of other communicative signals using the same modalities (speech and head). Prominence detection can be improved by the use of syllable-based analysis, which can be computed without linguistic information. Another important issue is to deal with the direction of gaze. This communicative signal conveys useful information during interaction and automatic analysis (human) and generation (robot/agent) should be investigated.

## Acknowledgments

We are grateful to Elisabetta Bevacqua for her advice in the organization of our work and her help on interfacing our software with GRETA.

We also want to acknowledge Yannis Stylianou for the feedback he gave during discussions on our project.

Some members of this were partly funded by Région Wallonne, in the framework of the NUMEDIART research programme.

FP6 IP CALLAS is also gratefully acknowledged for having funded participants of this project.

## REFERENCES

[1] K. Dautenhahn (2007) Methodology and Themes of Human-Robot Interaction: A GrowingResearch Field. International Journal of Advanced Robotic Systems 4(1) pp. 103-108.

[2] J.C. Baillie (2006) URBI tutorial [online] http://www.gostai.com/doc/en/urbi-tutorial/

[3] Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. (eds). Embodied Conversational Agents. MIT Press, 2000.

[4] E. Bevacqua, M. Mancini, and C. Pelachaud, A listening agent exhibiting variable behaviour, Intelligent Virtual Agents, IVA'08, Tokyo, September 2008.

[5] H. Vilhjalmsson, N. Cantelmo, J. Cassell, N. E. Chafai, M. Kipp, S. Kopp, M. Mancini, S. Marsella, A. N. Marshall, C. Pelachaud, Z. Ruttkay, K. R. Thorisson, H. van Welbergen, R. van der Werf, The Behavior Markup Language: Recent Developments and Challenges, Intelligent Virtual Agents, IVA'07, Paris, September 2007.

[6] T. Fong, I. Nourbakhsh and K. Dautenhahn (2003): A Survey of Socially Interactive Robots, Robotics and Autonomous Systems 42(3-4), 143-166.

[7] C. Breazeal (2004). "Social Interactions in HRI: The Robot View," R. Murphy and E. Rogers (eds.), in IEEE SMC Transactions, Part C.

[8] Oudeyer P-Y, Kaplan , F. and Hafner, V. (2007) Intrinsic Motivation Systems for Autonomous Mental Development, IEEE Transactions on Evolutionary Computation, 11(2), pp. 265--286.

[9] Sidner, C.L.; Lee, C.; Kidd, C.D.; Lesh, N.; Rich, C., "Explorations in Engagement for Humans and Robots", Artificial Intelligence, May 2005

[10] Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H. and Yan, H. (1999). "Embodiment in Conversational Interfaces: Rea." Proceedings of the CHI'99 Conference, pp. 520-527. Pittsburgh, PA.

[11] J. Rickel and W.L. Johnson, Animated Agents for Procedural Training in Virtual Reality: Perception, Cognition, and Motor Control, Applied Artificial Intelligence, 13, pp. 343-382, 1999.

[12] S. Kopp and I. Wachsmuth, Synthesizing Multimodal Utterances for Conversational Agents, The Journal Computer Animation and Virtual Worlds, 15(1), PP 39-52, 2004.

[13] J. Cassell and K. Thórisson, The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents, Applied Artificial Intelligence, 13(3), 1999.

[14] S. Marsella, Interactive Pedagogical Drama: Carmen's Bright IDEAS Assessed. Intelligent Virtual Agents, pp 1-4, 2003.

[15] E. André, T. Rist, S. van Mulken, M. Klesen and S. Baldes, The automated design of believable dialogues for animated presentation teams, in J. Cassell, J. Sullivan, S. Prevost and E. Churchill (Eds) Embodied Conversational Characters, MITpress Cambridge, MA, 2000.

[16] W.L. Johnson, H. Vilhjalmsson, S. Marsella, Serious Games for Language Learning: How Much Game, How Much AI?, 12th International Conference on Artificial Intelligence in Education, Amsterdam, The Netherlands, July 18-22, 2005.

[17] A.L. Baylor, S. Kim, C. Son and M. Lee, Designing effective nonverbal communication for pedagogical agents, Proceedings of AI-ED (Artificial Intelligence in Education), Amsterdam, July 2005.

[18] J. Gustafson, N. Lindberg, M. Lundeberg, The August sopken dialog system, Proceedings of Eurospeech'99, Budapest, Hungary, 1999.

[19] Martin, J.C., Buisine, S., Pitel, G., and Bernsen, N. O.: Fusion of Children's Speech and 2D Gestures when Conversing with 3D Characters. In T. Dutoit, L. Nigay and M. Schnaider (Eds.): Multimodal Human-Computer Interfaces. Special Issue of Signal Processing (Elsevier) Vol. 86, Issue 12, December 2006

[20] T. Bickmore, J. Cassell, Social Dialogue with Embodied Conversational Agents", in J. van Kuppevelt, L. Dybkjaer, N. Bernsen (Eds) Advances in Natural, Multimodal Dialogue Systems, Kluwer Academic, New-York, 2005.

[21] L. Hall, M. Vala, M. Hall, M. Webster, S. Woods, A. Gordon and R. Aylett, FearNot's appearance: Reflecting Children's Expectations and Perspectives, Intelligent Virtual Agents IVA, Marina del Rey, USA, pp. 407-419, 2006.

[22] M. Lohse, K. J. Rohlfing, B. Wrede; G. Sagerer, "Try something else!" - When users change their discursive behavior in human-robot interaction, IEEE Conference on Robotics and Automation, Pasadena, CA, USA, 3481-3486, 2008.

[23] J. Allwood, J. Nivre, and E. Ahlsen. On the semantics and pragmatics of linguistic feedback. Semantics, 9(1), 1993.

[24] I. Poggi. Backchannel: from humans to embodied agents. In AISB. University of Hertfordshire, Hatfield, UK, 2005.

[25] D. McNeil, Hand and mind: What gestures reveal about thought, Chicago IL, The University, 1992.

[26] Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta and Patrizia Paggio. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. Journal of Language Resources and Evaluation, Springer. Volume 41, 2007. pages 273-287

[27] Michael Kipp (2001) Anvil - A Generic Annotation Tool for Multimodal Dialogue. *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1367-1370.

[28] R. L. Brennan, D. J. Prediger: Coefficient κ: Some uses, misuses, and alternatives. In: Educational and Psychological Measurement. 41, 1981, 687–699.

[29] R. M. Maatman, Jonathan Gratch, Stacy Marsella: Natural Behavior of a Listening Agent. IVA 2005: 25-36.
[30] Pure Data: http://www.puredata.org
[31] The CMU Sphinx open source speech recognizer http://cmusphinx.sourceforge.net
[32] de Cheveigne, A., Kawahara, H.: YIN, a fundamental frequency estimator for speech and music. The Journal of the Acoustic Society of the America 111. 2002.
[33] Paul Taylor. The Tilt Intonation model, ICSLP 98, Sydney, Australia. 1982.
[34] B.M. Streefkerk, L. C. W. Pols, L. ten Bosch, Acoustical features as predictors for prominence in read aloud Dutch sentences used in ANNs, Proc. Eurospeech'99, Vol. 1, Budapest, 551-554, 1999.
[35] J.M.B. Terken, Fundamental frequency and perceived prominence of accented syllables. Journal of the Acoustical Society of America, 95(6), 3662-3665, 1994.
[36] N. Obin, X. Rodet, A. Lacheret-Dujour, "French prominence: a probabilistic framework", in International Conference on Acoustics, Speech, and Signal Processing (ICASSP'08), Las Vegas, U.S.A, 2008.
[37] V. K. R. Sridhar, A. Nenkova, S. Narayanan, D. Jurafsky, Detecting prominence in conversational speech: pitch accent, givenness and focus. In Proceedings of Speech Prosody, Campinas, Brazil. 380-388, 2008.
[38] D. Wang, S. Narayanan, An Acoustic Measure for Word Prominence in Spontaneous Speech. IEEE Transactions on Audio, Speech, and Language Processing, Volume 15, Issue 2, 690-701, 2007.
[39] P.Viola and M.J.Jones, "Robust Real-Time Face Detection", *InternationalJournal of Computer Vision*, 2004, pp 137-154.
[40] C.G. Harris and M.J. Stephens, "A combined corner and edge detector", Proc. Fourth Alvey Vision Conf., Manchester, pp 147-151, 1988
[41] Lucas, B., and Kanade, T. An Iterative Image Registration Technique with an Application to Stereo Vision, Proc. of 7th International Joint Conference on Artificial Intelligence (IJCAI), pp. 674-679.
[42] S. Young: "ATK: An application Toolkit for HTK, version 1.4, Cambridge University, May 2004
[43] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals. WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition. In Proc IEEE ICASSP, pages 81-84, Detroit, 1995.
[44] R. M. Maatman, Jonathan Gratch, Stacy Marsella, Natural Behavior of a Listening Agent. Intelligent Virtual Agents, IVA'05, 25-36, 2005.
[45] N. Ward, W. Tsukahara, Prosodic features which cue back-channel responses in English and Japanese. Journal of Pragmatics, 23, 1177-1207, 2000.
[46] Dehn, D. M. and van Mulken, S. (2000). "The impact of animated interface agents: a review of empirical research." International Journal of Human-Computer Studies(52): 1-22.
[47] Ruttkay, Z. and Pelachaud, C. (2004). From Brows to Trust - Evaluating Embodied Conversational Agents, Kluwer. http://wwwhome.cs.utwente.nl/~zsofi/KluwerBook.htm
[48] Buisine, S., Martin, J.-C. (2007) The effects of speech-gesture co-operation in animated agents' behaviour in multimedia presentations. International Journal "Interacting with Computers: The interdisciplinary journal of Human-Computer Interaction". Elsevier, 19: 484-493.
[49] Dan R. Olsen, Michael A. Goodrich (2003) Metrics for Evaluating Human-Robot Interactions. Performance Metrics for Intelligent Systems Workshop held in Gaithersburg, MD on September 16-18, 2003. http://www.isd.mel.nist.gov/research_areas/research_engineering/Performance_Metrics/PerMIS_2003/Proceedings/Olsen.pdf
[50] Mohan Rajesh Elara, Carlos A. Acosta Calderon, Changjiu Zhou, Pik Kong Yue, Lingyun Hu, Using Heuristic Evaluation for Human-Humanoid Robot Interaction in the Soccer Robotics Domain. Second Workshop on Humanoid Soccer Robots @ 2007 IEEE-RAS International Conference on Humanoid Robots Pittsburgh (USA), November 29, 2007 http://www.informatik.uni-freiburg.de/~rc06hl/ws07/papers/HSR-2-110.pdf

## X. Appendix

### Instructions

You are going to watch a short cartoon sequence.
Then you will have to tell this story.
Your story will be listened to and watched by the Greta virtual agent displayed on a screen and an Aibo robot at the same time.
Both will react to the story that you are telling.
Please tell the story as you would tell a story to two people listening to you at the same time.

### Evaluation form

Subject number :
Date of session :

*INFORMATION ABOUT SUBJECT*

Last name :
First name :
Age :
Male / Female
e-mail :

*GENERAL COMPARISON BETWEEN THE TWO LISTENERS*

Did you like telling your story to Greta?
I liked very much / I liked / I did not like
Did you like telling your story to Aibo?

I like very much / I liked / I did not like
Which one did you prefer?
Greta / Aibo
Why?
Which listener was mostly interested in your story?
Greta / Aibo
Which listener understood better your story?
Greta / Aibo
Which listener liked better your story?
Greta / Aibo
What did you like in Greta that was not present in Aibo?
What did you like in Aibo that was not present in Greta?

*EVALUATION OF FEEDBACK QUALITY*

Who most clearly displayed when it was interested in your story?
Aibo / Greta
Who most clearly displayed when it understood what you said?
Greta / Aibo
Who most clearly displayed when it liked what you said?
Aibo / Greta
How would you qualify the behavior displayed by Greta when you told your story ?
How would you qualify the behavior displayed by Aibo when you told your story ?

*PERCEPTION OF FEEDBACK SIGNALS*

How did Greta displayed that she was interested in your story?
How did Greta displayed that she was understanding in your story?
How did Greta displayed that she liked in your story?
How did Aibo displayed that it was interested in your story?
How did Aibo displayed that it was understanding in your story?
How did Aibo displayed that it liked in your story?

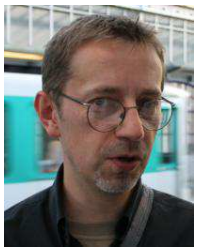
*GENERAL COMMENTS*

How did you evaluate the fact to interact with both at the same time?
Please feel free to provide any additional comments:

**Sàmer Al Moubayed** is a PhD student since 2008 at the Center of Speech Technology CTT, Royal Institute of Technology KTH, Sweden, and at the Graduate School of Language Technology GSLT, Gothenburg, Sweden. He obtained his MSc degree in Artificial Intelligence and Speech and Language Technology from KULeuven, Belgium.
His main research is carried out in the field of speech communication, and is doing his thesis in Multimodal Speech Synthesis, and is involved in many National and European projects. Samer main research interest is in Talking Agents and behavior prediction, machine learning and pattern recognition in speech, and psycholinguistics

**Malek Baklouti** graduated as engineer in from the Tunisian Polytechnic School and received the M.S. degree in Applied Mathematics in 2006. She is currently a PhD student in Robotics and Signal processing at Thalès Security System and Services, France and University of Versailles.
She will be in PAMI lab at the University of Waterloo (Canada) as a Visiting Scholar.

**Mohamed Chetouani** received the M.S. degree in Robotics and Intelligent Systems from the University Pierre and Marie Curie (UPMC), Paris, 2001. He received the PhD degree in Speech Signal Processing from the same university in 2004. In 2005, he was an invited Visiting Research Fellow at the Department of Computer Science and Mathematics of the University of Stirling (UK). Dr. Chetouani was also an invited researcher at the Signal Processing Group of Escola Universitaria Politecnica de Mataro, Barcelona (Spain).
He is currently an Associate Professor in Signal Processing and Pattern Recognition at the University Pierre et Marie Curie. His research activities, carried out at the Institute of Intelligent Systems and Robotics, cover the areas of non-linear speech processing, feature extraction and pattern classification for speech, speaker and language recognition.
He is a member of different scientific societies (ISCA, AFCP, ISIS). He has also served as chairman, reviewer and member of scientific committees of several journals, conferences and workshops.



**Thierry Dutoit** graduated as an electrical engineer and Ph.D. in 1988 and 1993 from the Faculté Polytechnique de Mons, Belgium, where he is now a full professor.
He spent 16 months as a consultant for AT&T Labs Research in Murray Hill and Florham Park, NJ, from July, 1996 to September, 1998. He is the author of two books on speech processing and text-to-speech synthesis, and the coordinator of the MBROLA project for free multilingual speech synthesis.
T. Dutoit was an Associate Editor of the IEEE Transactions on Speech and Audio Processing (2004-2006) and a member of the INTERSPEECH'07 organization committee. He was the initiator of eNTERFACE workshops and the organizer of eNTERFACE'05.

**Ammar Mahdhaoui** was born in Tunisia, on january 31, 1983. He received the M.S. degree in Engineering of Person-System communication from the university of Grenoble, Grenoble, 2007. Since October 2007, he is PHD Student in signal processing and pattern recognition at the University Pierre and Marie Curie Paris-6, his research activities carried out at the Institute of Intelligent Systems and Robotics.

**Jean-Claude Martin** is Associate Professor at CNRS-LIMSI, France.
His research topic is multimodal communication, both in human-human and human-computer contexts ; the study of individual differences, the multimodal expression and perception of social behaviors and the evaluation of user's multimodal interaction with embodied conversational agents.
He received the PhD degree in Computer Science in 1995 from the Ecole Nationale Supérieure des Télécommunications (ENST, Paris). He passed his habilitation to direct research in 2006 on Multimodal Human-Computer Interfaces and Individual Differences.
Annotation, perception,
representation and generation of situated multimodal behaviors. He is the head of the Conversational Agents topic of research within the Architecture and Models for Interaction Group (AMI). He co-organised a series of three international workshops on multimodal corpora at LREC 2002, 2004, 2006 and he is a guest editor of a special issue of the international Journal on Language Resources and Evaluation to appear in 2007 on multimodal corpora.

**Stanislav Ondas** graduated from the Technical University of Košice in 1999. In 2007 he finished your PhD. study at the Department of Electronics and Multimedia Communications at the same university and he is currently an assistent at mentioned department.
His research interests in spoken dialogue systems, dialogue management, voice service designing and natural language processing**.** He has been involved in several national projects related to spoken and multimodel dialogue systems (ISCI, MOBILTEL).

From October 2008, **Catherine Pelachaud** will be a director of research at CNRS in LTCI, TELECOM ParisTech**.**She received her PhD in Computer Graphics at the University of Pennsylvania, Philadelphia, USA in 1991. Her research interest includes representation language for agent, embodied conversational agent, nonverbal communication (face, gaze, and gesture), expressive behaviors and multimodal interfaces. She has been involved in several European and national projects related to multimodal communication (EAGLES, IST-ISLE), emotion (FP5 NoE Humaine, FP6 IP CALLAS, FP7 STREP SEMAINE), to believable embodied conversational agents (IST-MagiCster, FP5 PF-STAR, RIAM ACE, ANR My-Blog-3D), and to handicap (CNRS-Robea HuGeX, RIAM LABIAO).

**Jérôme Urbain** graduated as an electrical engineer from the Faculté Polytechnique de Mons (FPMs), Belgium, in 2006. He is currently PhD student at the Signal Processing and Circuit Theory (TCTS) Lab of the same University, working on speech processing in the framework of FP6 IP CALLAS.
He is focusing on laughter modeling, synthesis and recognition.

**Mehmet Yilmaz** is a senior undergraduate student at Koç University, Electrical and Electronics Department. His research interests are visual tracking, and interactive multimodal systems.